John von Neumann Institute for Computing

Study of Protein Structural Descriptors: Towards Similarity and Classification

P. Jain, J. D. Hirst

published in

From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **36**, ISBN 978-3-9810843-2-0, pp. 165-167, 2007.

http://www.fz-juelich.de/nic-series/volume36

Study of Protein Structural Descriptors: Towards Similarity and Classification

Pooja Jain and Jonathan D. Hirst

School of Chemistry,
University of Nottingham, Nottingham,
NG7 2RD, United Kingdom
E-mail: {*pcxpj1, jonathan.hirst*}*@nottingham.ac.uk*

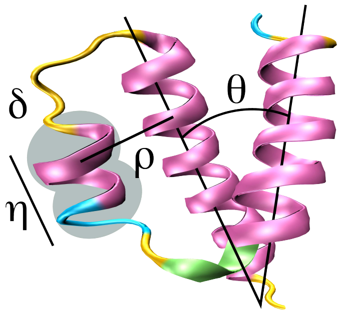
We have investigated structural descriptors for structural similarity and classification of 77 proteins extracted from SCOP. A Support Vector Machine was trained to predict structural similarity based on paired protein profiles, composed of structural descriptors derived from the geometric properties of secondary structure elements. Ten fold cross-validation, against the standard similarity measure from DALI gave a cross-validated correlation coefficient, q^2, of 0.91. A coefficient of dissimilarity was derived as the Euclidean distance among different descriptor types of two proteins. This coefficient was evaluated for the classification of protein pairs to different levels in SCOP hierarchy.

1 Introduction

Protein structure comparison can provide useful information on the biological function of a protein[1] and can imply evolutionary relationships between proteins with low sequence similarity. This information is crucial in the identification of new protein folds and understanding the organisation of the known universe of protein structures. The aim of this work is to perform protein structure analysis and comparison through the use of structural descriptors. These are numerical values that characterise the secondary structure elements of a protein. For example, they may represent the physico-chemical properties or geometric properties of the secondary structure elements derived from the 3D coordinates. In our study we make use of SCOP, DALI and USM. SCOP is a curated database which aims to provide a comprehensive description of the structural and evolutionary relationships between all protein structures[2]. The principal levels in the SCOP hierarchy are class, fold, superfamily and family. DALI is a common and popular structural alignment and comparison method[1]. It represents a protein as a matrix of contact patterns between successive hexapeptide fragments and makes comparisons with such matrices of other proteins. USM is based on the comparison of compressed protein contact maps using the principle of Kolmogorov complexity[3].

2 Structural Descriptors

A protein was defined in terms of structural descriptors derived from its secondary structure elements. A set of 77 proteins containing exactly and only three α helices from the all-α class of SCOP database was used. The descriptors, such as the pairwise separation ρ between the centre of mass COM of any two secondary structure elements i and j of lengths n and m in a protein along Cartesian coordinates A, the relative orientation cosθ,



$$\rho_{ij} = \sqrt{\sum_{A=X,Y,Z}(COM_{Ai} - COM_{Aj})^2}$$

$$cos\theta_{ij} = \frac{\vec{V_i} \cdot \vec{V_j}}{|V_i||V_j|}$$

$$\delta_i = \frac{\sum_1^n \delta_n}{n}$$

$$\eta_i = n_i$$

Figure 1. A pictorial representation of structural descriptors and their mathematical definitions.

the individual surface accessibility δ and the length η for each of them were derived from the DSSP assignments[4]. Figure 1 illustrates these descriptors in a protein structure and give their mathematical definitions.

3 The Protein Profile

The geometric profiles made up of above defined descriptors was used to pair up any two proteins. Such paired protein profiles included along with the 12 descriptors for each of the proteins, the RMS difference between respective descriptor types for that pair. For the 77 proteins, 2,926 paired protein profiles containing 28 elements each were generated.

3.1 Profile Based Structural Similarity

Using Support Vector Machines (SVMs) the paired protein profiles were subjected to the non-linear (multivariate) regression against the protein similarity values assigned by DALI and USM. The model was trained by the Sequential Minimal Optimisation algorithm for regression analysis (SMOreg)[5] from Waikato Environment for Knowledge Analysis (WEKA) software package[6]. Parameter tuning was performed to choose the best values for complexity parameter and kernel function. Finally, the model assessment was performed using 10-fold cross-validation.

3.2 Profile Based Structural Classification

For a pair of proteins x and y, a coefficient of dissimilarity Ω_{xy} that gives the Euclidean distance between them was derived from the RMS difference of different descriptor types as below:

$$\Omega_{xy} = \sqrt{\rho_{xy}^{rmsd} + \theta_{xy}^{rmsd} + \eta_{xy}^{rmsd} + \delta_{xy}^{rmsd}}$$

The higher Ω more dissimilar are the proteins.

4 Results

The outcome of multivariate regression of paired protein profiles against the similarity values assigned by USM and DALI was significant with cross-validated correlation coefficients (q^2) of 0.74 and 0.91, respectively.

The structural classification of protein pairs was based on the coefficient of dissimilarity Ω. The protein pairs belonging to the same family congregated towards a lower dissimilarity threshold, whilst those sharing the same fold were associated with higher values.

5 Concluding Remarks

The results from multivariate regression of protein profiles suggest their potential as a representation of protein structures and there further use in the protein structure comparison. Ideally, the profile based structural classification of proteins to different levels in SCOP hierarchy should be distinctive. Our results show some bias towards this. Efforts to improve this are ongoing. Analysis continues on larger datasets comprising proteins containing three or four secondary structure elements.

Acknowledgements

This work was supported by BIOPTRAIN project MEST-CT-2004-007597 under the Sixth framework program of the European Community. We thank the University of Nottingham for providing high performance computational resources.

References

1. L. Holm and C. Sander, *Protein-structure comparison by alignment of distance matrices*, J. Mol. Biol. **233**, 123–138, 1993.
2. A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, *SCOP: a structural classification of proteins database for the investigation of sequences and structures*, J. Mol. Biol. **247**, 536–340, 1995.
3. N. Krasnogor and D. A. Pelta, *Measuring the similarity of protein structures by means of the universal similarity metric*, Bioinformatics **20**, 1015–1021, 2004.
4. W. Kabsch and C. Sander, *Dictionary of Protein Secondary Structure: Pattern recognition of Hydrogen-bonded and Geometrical Features*, Biopolymers **22**, 2577–2637, 1983.
5. G. W. Flake and S. R. Lawrence, *Efficient SVM regression training with SMO*, Machine Learning **46**, 271–290, 2002.
6. I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2nd Edition, 2005.